Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. Filip Rafaj
Název práce	Pojmenované entity a ontologie metodami hlubokého učení
Rok odevzdání	2021
Studijní program	Informatika **Studijní obor** Umělá inteligence

Autor posudku	Jan Hajič **Role** Vedoucí
Pracoviště	ÚFAL

Text posudku:

Diplomová práce "Pojmenované entity a ontologie metodami hlubokého učení" (Named Entity Linking by Deep Learning) se zabývá dvojicí navazujících úloh, tzv. Named Entity Recognition (NER) a Named Entity Disambiguation (NED), které dohromady tvoří tzv. Named Entity Linking (NEL). Cílem této úlohy je identifikovat úseky textu, které odkazují na individuální "hesla" v nějaké(ých) ontologii(ích) neboli "bázích znalostí (o světě)" (Knowledge Bases). V konkrétním případ této DP byla použita Wikipedie. Úloha byla řešena pomocí částečně řízených metod (semi-supervised learning) hlubokého učení (Deep Learning pomocí DNN), a to za použití moderních postupů (vč. kontextových embedingů).

Úvod práce popisuje úlohu a její nastavení, uvádí obecné práce v oblasti hlubokého učení, ze kterých vychází a popisuje strukturu práce, která má tři kapitoly. V první kapitole popisuje znovu podrobněji úlohu, kterou řešil, včetně příkladů, základní použité architektury DNN, a Word Embeddings (od standardních, včetně prací, které tento směr naznačily již v 70. letech 20. století, až po kontextové, popsané v používané verzi poprvé v letech 2016-8). Dále zde popisuje použitá data a standardní evaluační metriky pro danou úlohu. Na závěr první kapitoly je souhrn prací, které se v poslední době zabývaly podobnou tématikou (pro NER, NED i NEL, včetně "joint" modelování NEL), včetně odkazu na SoA výsledky na autorem použitém datasetu (CoNLL-YAGO).

Ve druhé kapitole popisuje autor jím vytvořené a použité metody, architekturu modelů, způsob použití různě strukturovaných embedingů, od dotrénovaných BERT kontextových embedingů přes embedingy pro úseky textu jako kandidáty pro linking až po embedingy pro entity vytvořené na základě textů popisujících jednotlivé entity (zde tedy texty z Wikipedie). Je zde popsáno i "skórování" modelu (lokální i globální), ztrátová funkce pro trénování a způsob inference ("runtime" pro vlastní určování NEL na neznámém textu).

Třetí kapitola popisuje provedené experimenty a jejich výsledky na datové sadě CoNLL-YAGO a dalších datových sadách. Vyhodnocení je provedeno jak tzv. silnou, tak slabou metrikou (slabá metrika nevyžaduje přesnou korespondenci správného úseku textu).

Závěr práce pak popisuje východiska, provedené experimenty a shrnuje výsledky: základ, ze kterého experimenty s různými typy kontextových embedingů vycházely, tj. práce Kolitsas et al. (2018), byl zlepšen o 1-2 procentní body. Některé pozdější práce dávají v silné metrice dnes lepší výsledky, ale vyžadují dotrénování BERTu (fine-tuning). Paralelně s touto DP vyšla práce, která také používá BERT, což lze považovat za potvrzení správného směru zvoleného v této DP. Kód práce je volně k dispozici na Githubu.

Práce je psaná velmi dobrou angličtinou s minimem chyb (spíše neobratností než vyložených chyb). Je stručná, ale všechny nutné informace (např. pro replikovatelnost výsledků) jsou v ní (kap. 2 a 3) obsaženy. Používané pojmy, včetně speciálních a autorem definovaných, jsou na začátku kapitol vysvětleny. Množství experimentů je pro DP více než dostačující.
Vyzdvihnout je třeba i fakt, že práce dosahuje výsledků blízkých SoA (bráno absolutně, tj. ve světě v r. 2020), a při použití slabé metriky je překonává (i když zde lze přímo porovnávat jen se starší prací z r. 2018). Literatura je adekvátní a více než dostatečná. Při obhajobě by bylo vhodné odpovědět na otázku, zda (případně jak) by autor ošetřil případy, kdy entita v použité ontologii neexistuje (a tedy nebylo možné natrénovat embedingy pro entity), a zda je v daném řešení vůbec možné, příp. zda by bylo možné řešení nějak upravit alespoň pro označování NER (bez disambiguace).

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 26. January 2021

Podpis

Prof. RNDr. JAN HAJIČ, Dr.

Digitally signed by Prof. RNDr. JAN HAJIČ, Dr.
Date: 2021.01.28 14:07:45 +01'00'